International Journal of Advanced Intelligence
Volume 0, Number 0, pp.XXX-YYY, November, 20XX.

AIA International Advanced Information Institute

Deduction of Translation Relations between New Short Sentences in Chinese and Japanese Using Analogical Associations

Wei Yang
Graduate School of IPS, Waseda University, 2-7 Hibikino, Wakamatsu, Kitakyushu Fukuoka, 808-0135, Japan
kevinyoogi@akane.waseda.jp

Hao Wang
Graduate School of IPS, Waseda University, 2-7 Hibikino, Wakamatsu, Kitakyushu Fukuoka, 808-0135, Japan
oko_ips@ruri.waseda.jp

Yves Lepage
Graduate School of IPS, Waseda University, 2-7 Hibikino, Wakamatsu, Kitakyushu Fukuoka, 808-0135, Japan
yves.lepage@waseda.jp

Received (September 2013)
Revised (January 2014)

Bilingual parallel corpora are an essential resource in machine translation (MT). There already exist freely available corpora for European languages, but almost none between Chinese and Japanese. We propose to construct a free Chinese-Japanese quasi-parallel corpus by using analogical associations based on linguistic examples collected from the Web. We first over-generate new candidate sentences by analogy. Then we filter them by attested N-sequences and obtain new sentences at least 99% correct on the grammatical level. We then deduce translation relations across languages based on similarity computation and obtain thousands of quasi-parallel Chinese-Japanese sentence pairs with their associated similarity scores from several tens of thousands sentences.

Keywords: Quasi-parallel Corpus; Analogies; Clustering; Machine Translation.

1. Introduction

Recent corpus-based approaches to statistical or example-based machine translation make use of large amounts of aligned sentences as training data. There already exist freely and easily available corpora for European languages, such as the Europarl parallel corpus[1] which is a collection of parallel text in 11 offical languages of the European Parliament, or the DGT-TM, a freely available translation memory in 22 languages[2], but almost none between Chinese and Japanese publicly available parallel corpora for the development of language technology.

Some research institutions have tried to construct Chinese-Japanese bilingual parallel corpora, e.g., NICT (National Institute of Information and Communica-

tions Technology, Japan.), Kyoto U. (Kyoto University, Japan.) and HIT (Harbin Institute of Technology, China.).

(i) NICT created a Japanese-Chinese corpus of 38,383 sentences by selecting Japanese sentences from the Mainichi Newspaper and translating them manually into Chinese. They then annotated the corpus with morphological and syntactic structures and alignments at word and phrase levels[3].
(ii) Kurohashi-Kawahara Lab[a] in Kyoto U. created the Japanese-English-Chinese (JEC) Basic Sentence Data based on the Japanese Basic Sentence Data, automatically extracted from the Kyoto University Case Frame data. Their data contain manually modified 5,304 short sentences, and then manually translated data from Japanese into English and Chinese as a NICT MASTAR Project in Multilingual Translation Laboratory[b].
(iii) HIT, constructed the Olympic Oriented Chinese-English-Japanese Trilingual Corpus[4] from a Chinese-English parallel corpus collection by adding Japanese translations, for the development of Natural Language Processing (NLP) for Beijing 2008 Olympic Games in multi-domain. The resource consists of 54,043 sentences pairs.

Such corpora are all translated from one language into another language manually. None of them has been constructed automatically. Except for the JEC Basic Sentence Data, all the rest is not publicly or freely available, due to copyright problems. These parallel corpora are small in comparison to the above-mentioned multilingual corpora in European languages.

The constitution of large collections of aligned sentences is a problem for less documented language pairs. But, it is to be noticed that a less documented language pair may involve two well-documented languages, as is the case for the languages we address here: Chinese and Japanese.

Recognizing associations between words or sentences is an important task in Natural Language Processing (NLP). Recently, a number of works in NLP make use of corpus-based analogical techniques. Lepage and Denoual[5] (2005) use proportional analogies to translate sentences for machine translation; Turney and Littman[6] (2005) show the use of different machine techniques to answer SAT tests (analogical puzzles on words); Stroppa and Yvon[7] (2005) show the application of analogy in morphological analysis, they reported state-of-the-art results for three languages; Lavallée and Langlais[8] (2009) also work on morphology acquisition, they show their attempts in designing practical systems based on the analogical principle. In all these works, associations are obtained using analogical relations [*hand : glove :: foot : shoe*] or [*to create : creator :: to translate : translator*] or for sentences

[a]Kurohashi-Kawahara Lab: `http://nlp.ist.i.kyoto-u.ac.jp`
[b]Multilingual Translation Laboratory: `http://www.nict.go.jp/en/univ-com/multi_trans/`

Do you like music? : *Do you go to concerts?* :: *Do you like classical music?* : *Do you go to classical concerts?*

In this paper, we propose to construct a *quasi-parallel* Chinese-Japanese corpus by making use of such analogical associations based on Chinese and Japanese linguistic resources collected from the Web using an in-house Web-crawler. A "*quasi-parallel corpus*" is different from a parallel corpus in that it contains sentences that are nearly the exact translation to each other. To characterize the degree of translation, similarity scores between the sentences are provided for each sentence pair.

We propose to cluster large amounts of Chinese and Japanese short sentences using analogical associations. Such clusters can be considered as rewriting models that can generate new sentences. For filtering over-generated sentences and enforce fluency of expression and adequacy of meaning, we use attested N-sequences method. We also compute the similarity between the clusters across languages using a classical Dice formula. Based on the similarity between the clusters across languages, and the similarity between sentences used for new sentences generation, we can deduce translation relations between newly generated sentences. The set of such newly created sentences, with their translation scores, will constitute a quasi-parallel Chinese-Japanese corpus without copyright problems, as all the sentences contained will have been created by our programs.

2. The Chinese and Japanese Linguistic Resources Used

For our experiments, we use sentences of less than 30 characters in size. These sentences have been collected from the Web using an in-house Web-crawler. The use of the Web should ensure that our data is made of natural sentences. The main websites from which we collected monolingual sentences are "Yahoo China[c]", "Yahoo China News[d]", "douban[e]" for Chinese and "Yahoo! JAPAN[f]", "Mainichi Japan[g]", "Rakuten Japan[h]" for Japanese. We cleaned up these data by filtering out any sentence containing undesirable characters or symbols. For Chinese data, we retained sentences that contain only simplified Chinese characters. Figure 1 illustrates the increasing tendency and the number of collected (raw) and retained (after filtering) Chinese and Japanese short sentences in one year from month to month.

Table 1 shows the statistics about these collected-filtered Chinese and Japanese short sentences. About half (52%) of the Chinese sentences and more than half (60.6%) of Japanese short sentences are kept after filtering. Table 2 shows samples

[c]Yahoo China: `http://cn.yahoo.com`
[d]Yahoo China News: `http://news.cn.yahoo.com`
[e]douban: `http://www.douban.com`
[f]Yahoo! JAPAN: `http://yahoo.co.jp`
[g]Mainichi Japan: `http://mainichi.jp`
[h]Rakuten Japan: `http://www.rakuten.co.jp`

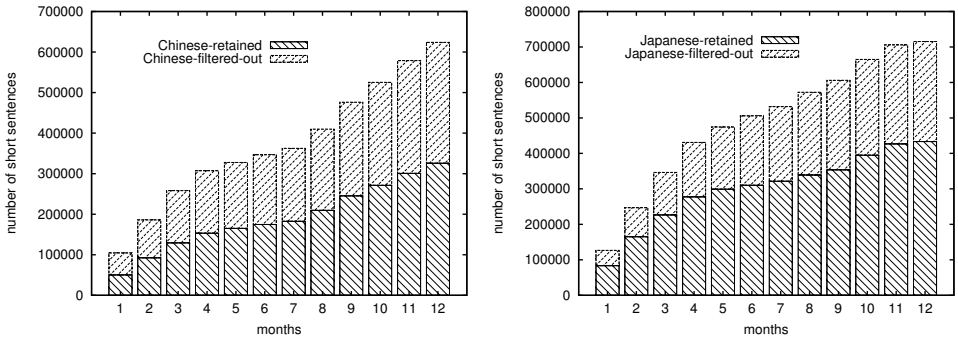

Fig. 1. Statistics about Chinese and Japanese raw-filtered short sentences collected from the Web in one year.

of raw-filtered sentences encountered in sentences filtering processing. The quality of these kept short sentences has been estimated by hand and is at least 98% correct sentences (p-value = 0.02).

Table 1. Statistics on the filtered Chinese and Japanese monolingual short sentences.

	# of different sentences (collected)	# of different sentences (filtered)	size of sentences in characters mean ± std.dev.	# of characters	# of words
Chinese	623,929	325,815	11.29 ± 7.24	3,609,708	2,445,764
Japanese	715,432	433,292	16.06 ± 7.43	7,053,924	4,116,804

We also collected and processed some Chinese-Japanese parallel data, as a part of our experimental data for assessment purposes:

- the JEC Basic Sentence Data (Kyoto U. and NICT, 2011) with 5,304 Chinese-Japanese sentence pairs.
- Chinese-Japanese Learning Corpus: from "日语学习网[i]" and "沪江网[j]", we obtained about 15,302 Chinese-Japanese parallel sentences after processing.

The total number of these Chinese-Japanese parallel corpora in lines is 20,606, including 16,259 different sentences for Chinese and 20,079 different sentences for Japanese. Table 3 gives the statistics on these Chinese-Japanese parallel sentences. Table 4 shows some samples of these parallel sentences. As an ideal configuration, here we suppose the similarity between each sentence pair is 1.000.

[i] 日语学习网 (Japanese Learning net): `http://jp.tingroom.com`
[j] 沪江网 (HuJiang): `http://www.hujiang.com`

Table 2. Samples of raw-filtered sentences encountered in sentences filtering processing. The filtered out sentences are struck through in the table.

	Raw-Filtered Sentences
Chinese	~~#侦探的理直气壮,医生的愠怒不屑。~~ 大庆职业技能鉴定在大庆炼化展开 ~~繁體字比簡體字難寫的多呀!~~ 一天早晨,三只熊打算早饭前一起去散散步。 05月11号 蜡笔小新第二部第62集 . . .
Japanese	まず楽曲が古くさくてセンスが無い。 2002年「リレキショ」で文藝賞を受賞しデビュー。 2008年にオープン。 ~~見どころ満載*です。~~ 新感覚ファンタジー、第四弾! 早稲田大学卒業後、慶応義塾大学大学院修了。 . . .

Table 3. Statistics on the Chinese-Japanese parallel corpora.

	Language	# of different sentences	size of sentences in characters mean ± std.dev.	# of characters	# of words
JEC	Chinese	5,299	12.31 ± 4.40	65,219	43,761
	Japanese	5,304	16.29 ± 5.38	86,409	53,654
日语学习网	Chinese	10,960	11.32 ± 4.46	125,862	86,997
	Japanese	14,775	17.22 ± 8.31	254,650	137,070

Table 4. Samples of Chinese-Japanese parallel corpus with their translation similarity score.

Chinese	**Japanese**	*Sim*
我想山田是受大家欢迎的那种人。	山田はみんなに好かれるタイプの人だと思う。	1.000
6年级学生练习了唱歌。	6年生が、歌の練習をしました。	1.000
以磨练和培养为目的。	錬磨及び育成を目的とします。	1.000
这种损人的行为应当受到惩罚。	はた迷惑な罰当たりな行為だ。	1.000
午休时间，在专业学校附近的便利店旁会聚集很多年轻人。	昼休み、専門学校の近所のコンビニに集う若者たち。	1.000
除了餐饮费之外，还要向您收取服务费。	飲食代とは別にサービス料をいただきます。	1.000

3. Building Analogical Clusters

3.1. *Proportional analogies*

Proportional analogies establish a general relationship between four objects, A, B, C and D. An analogy $A : B :: C : D$ states that 'A is to B as C is to D'. Previous

research by Lepage (1998) in Ref. 5 proposes an efficient algorithm for the resolution of analogical equations. The algorithm is based on the following formalisation of analogies, basically, in terms of similarity. More precisely, the formalization bases on counting numbers of occurrences of characters and the computing of edit distances between strings of characters. It is given by Formula (1).

$$A : B :: C : D \Rightarrow \begin{cases} |A|_a - |B|_a = |C|_a - |D|_a, \forall a \\ d(A,B) = d(C,D) \\ d(A,C) = d(B,D) \end{cases} \quad (1)$$

where $|A|_a$ stands for the number of occurrences of character a in string A and $d(A,B)$ stands for the edit distance between strings A and B with only insertion and deletion as edit operations. As B and C may be exchanged in an analogy, the constraint on edit distance has also to be verified for $A : C :: B : D$, i.e., $d(A,C) = d(B,D)$. A very efficient and fast way to compute the distance between two sentences seen as strings of characters is to compute their similarity using the fast bit string algorithm described by Allison and Dix (1986) in Ref. 6 and then derive the value of the canonical distance: $d(A,B) = |A| + |B| - 2 \times s(A,B)$.

In our research, we extract pairs of sentences that follow the above formula for proportional analogies. For instance, the two following pairs of Japanese sentences are said to form an analogy:

紅茶が飲みたい。 : あなたは紅茶が好きですか。 :: ビールが飲みたい。 : あなたはビールが好きですか。

I'd like a black tea. : *Do you like black tea?* :: *I'd like a beer.* : *Do you like beer?*

Because the relational similarity between the sentence pair on the left side of '::' is the same as between the sentence pair on the right side, we call any such two pairs of sentences sentential analogies (English part is translation of the example in Japanese).

When several sentential analogies involve the same pairs of sentences, they form a series of analogous sentences, and they can be written on a line like in:

紅茶が飲みたい。 : あなたは紅茶が好きですか。 :: ビールが飲みたい。 : あなたはビールが好きですか。 :: ジュースが飲みたい。 : あなたはジュースが好きですか。

I'd like a black tea. : *Do you like black tea?* :: *I'd like a beer.* : *Do you like beer?* :: *I'd like a juice.* : *Do you like juice?*

More conveniently, they can also be written on a sequence of several lines like:

紅茶が飲みたい。　：あなたは紅茶が好きですか。
ビールが飲みたい。　：あなたはビールが好きですか。
ジュースが飲みたい。　：あなたはジュースが好きですか。

Such a sequence of lines, where each line contains one sentence pair and where any two pairs of sentences form a sentential analogy, we shall call a analogical cluster. This is the case in the example above where all the three possible sentential analogies listed below hold:

紅茶が飲みたい。：あなたは紅茶が好きですか。::ビールが飲みたい。：あなたはビールが好きですか。

紅茶が飲みたい。：あなたは紅茶が好きですか。::ジュースが飲みたい。：あなたはジュースが好きですか。

ビールが飲みたい。：あなたはビールが好きですか。::ジュースが飲みたい。：あなたはジュースが好きですか。

To repeat in slightly different terms, using our formula definition of analogy, a cluster is a series of pairs of analogous sentences written over several lines, and any two pairs of analogous sentences form a sentential analogy. We call the size of a cluster the number of sentential pairs.

3.2. *Experiments in clustering short sentences in Chinese and Japanese*

We performed experiments based on proportional analogy with Chinese and Japanese monolingual data respectively. The number of lines of unique sentences used is 47,674 for Chinese and 95,130 for Japanese (including 16,259 Chinese unique sentences and 20,079 Japanese unique sentences of 20,606 Chinese-Japanese parallel sentences) from the collected data after filtering. Table 5 summarizes some statistics on the clusters produced on a machine with a 11.8Gb memory and 3.2GHz processor in a little bit more than 5 hours for Chinese and and 21 hours for Japanese. The clustering process built 28,455 Chinese clusters (37,185 for Japanese; in the sequel the figures in parentheses are for Japanese), which 2,893 (8,445) contain only two pairs of sentences (called small clusters). The remaining 25,562 (28,740) clusters contain more than 3 pairs of sentences (called large clusters). The larger a cluster, the more productive it is. Figure 2 (upper graphs) plots the number of analogous sentence pairs produced in clusters against the number of these clusters. We gave

an identifier to each cluster with a natural number. This identifier is its rank in the ordering by decreasing size (as shown in lower graphs of Figure 2).

Table 5. Statistics on the Chinese and Japanese clusters production.

		Chinese	**Japanese**
# of different sentences	(i)	47,674	95,130
# of clusters	(ii)	28,455	37,185
# of small clusters		2,893	8,445
# of large clusters		25,562	28,740
Time spent (h)		5.37	21.34
# of different sentences involoved	(iii)	11,517	9,374
average # of clusters per sentence (all sentences)	(ii)/(i)	0.60	0.39
average # of clusters per sentence (involved only)	(ii)/(iii)	2.48	3.61

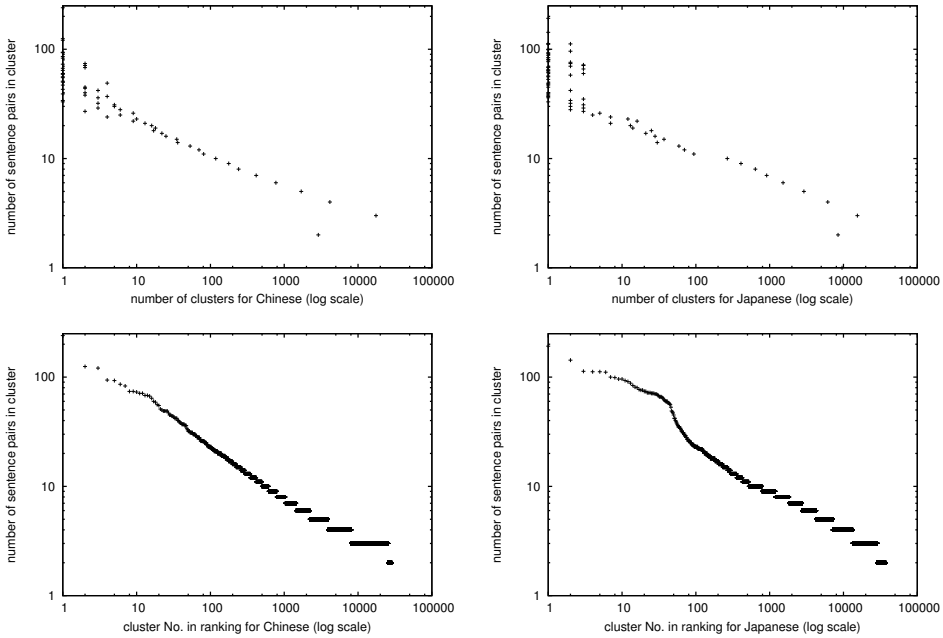

Fig. 2. Statistics on the number of analogous sentence pairs produced in clusters against the number of these clusters for Chinese (on the left) and Japanese (on the right) respectively (upper graphs), the other two graphs (lower graphs) shows the identifier assigned to a cluster will be a rank in the ordering by size for clusters. The lower the number the higher the ranking.

We give some examples of the clusters we constructed. The sentence pairs where on the left and right are not necessarily the paraphrases, i.e., in the similar meaning to each other. Because different clusters illustrate different linguistic or semantic features, the same sentence may appear in different clusters. For instance, the Japanese sentence 改善お願いします。 /kaizen onegai shimasu/ 'Improve it, please.' appears on the right in the cluster in Table 6 (indicated with a '▲'). The linguistic interpretation of this cluster is that the noun 検討 /kentou/ 'investigate' is exchanged with 改善 /kaizen/ 'improve' in similar situational and structural contexts in different meaning. The same sentence also appears in the cluster in Table 7. This cluster shows the insertion of the degree adverbial "よろしく" /yoroshiku/. In terms of linguistic feature, it lies between a neutral and a more polite form of expression.

We also found that the position of the changes in a cluster is not necessarily exactly the same. As shown in Table 8, obviously the position of insertion of the Chinese adverbial "非常" /fēicháng/ 'very much' in sentence pairs 1, 2, 4 is different from the sentence pairs observed in 3, and 5.

From the point of view of the size of the clusters, the largest cluster for Chinese in our experiment contains 240 pairs of sentences. The interpretation of this cluster is the insertion of the Chinese degree adverbial 很 /hěn/ 'very'. It is similar to the clusters we give in Table 8. The largest cluster for Japanese contains 192 pairs of sentences. This cluster exhibits similar phenomena as the cluster shown in Table 7. It lies between a neutral and a more polite form of speaking or expresses a solemn decision by adding a auxiliary verb です /desu/.

The next largest and the third largest clusters for Chinese contain 125 and 121 sentence pairs respectively, they both show the insertion model of Chinese word 的 /de/. They were separated into two clusters due to the difference of distances between the pairs of sentences on the left and right, they also reflect different linguistic phenomena. The next largest cluster (the distance is 5) shows subject-predicate phrase change to nominal endocentric phrase (Table 9). The third largest clusters (the distance is 1) reflect some kinds of usage of the word 的: (1) make a word or phrase into an adjective; (2) change a word or phrase into a demonstrative pronoun; (3) express the relationship between words (4) as the auxiliary word, that in the end of the sentence to strengthen a affirmative tone. All these usages are shown in Table 10.

A manual inspection of the other larger Japanese clusters obtained shows that the clusters illustrate a range of linguistic phenomena:

(i) Orthographical variations, mainly for Japanese with writing in kanji vs kana (e.g., (ja) 下さい /kudasai/ vs ください /kudasai/, they both mean 'please'.);
(ii) Exchange of place names, people names etc. (e.g., (ja) 秋田 /Akita/ and 福島 /Fukusima/.);
(iii) Some clusters contain dozens of pairs of sentences that illustrate the exchange of digits (e.g., (ja) 8月18日生まれ /hachi gatsu jyuhachi nichi umare/ 'Born on August 18th.' and 8月28日生まれ /hachi gatsu nijyuhachi nichi umare/ 'Born

Table 6. A Japanese cluster (identifier 3807) that illustrates the substitution of the verb “検討” /kentou/ ‘investigate’ with “改善” /kaizen/ ‘improve’.

検討願います。	:	改善願います。
検討をお願いします。	:	▲改善をお願いします。
検討よろしくおねがいします。	:	改善よろしくおねがいします。
検討よろしくお願いします。	:	改善よろしくお願いします。
ご検討ください。	:	ご改善ください。
検討お願いします 。	:	改善お願いします。

Table 7. A Japanese cluster (identifier 7441) that illustrates the possible insertion of the adverbial “よろしく” /yoroshiku/.

▲改善お願いします。	:	改善よろしくお願いします。
復旧作業お願いします。	:	復旧作業よろしくお願いします。
復旧をお願いします。	:	復旧をよろしくお願いします。
ご確認お願いし ます 。	:	ご確認よろしくお願いします。

Table 8. A Chinese cluster (identifier 515) that illustrates the insertion of the adverbial “非常” /fēicháng/ ‘very much’.

操作方便	:	操作非常方便
效果不错	:	效果非常不错
值得推荐	:	非常值得推荐
孩子喜欢	:	孩子非常喜欢
值得称赞	:	非常值得称赞
.....		.....

Table 9. A Chinese cluster (identifier 3) that illustrates the insertion of the word “的” /de/, shows the subject-predicate phrase change to nominal endocentric phrase.

画面漂亮	:	漂亮的画面
游戏很好玩	:	很好玩的游戏
故事有趣	:	有趣的故事
软件非常好用	:	非常好用的软件
节奏欢快	:	欢快的节奏
.....		.....

Table 10. A Chinese cluster (identifier 3) that illustrates the insertion of the word "的" /de/ reflect different linguistic phenomenas.

挺简单	:	挺简单的	(1)
没声音	:	没声音的	(1)
其他	:	其他的	(2)
唱歌	:	唱歌的	(2)
他评论	:	他的评论	(3)
明明是免费	:	明明是免费的	(4)
.....		.....	

on August 28th.').

(iv) Change of attributive or adverbial to other expressions (e.g., (ja) change adverbial 超 /chou/ to とても /totemo/, they both mean 'very much');

(v) etc.

4. Generation of New Sentences by Using Analogical Associations

4.1. *Generation of new sentences*

We now show how to generate new sentences based on analogical relations. Following insights by Saussure in Ref. 7 we use analogy as a synchronic operation by which, given two related forms and only one form, the fourth missing form is coined. Applied on sentences, this principle can be illustrated as follows:

紅茶が飲みたい。 : あなたは紅茶が好きですか。 :: ビールが飲みたい。 : x
$\Rightarrow$ $x =$ あなたはビールが好きですか。

If the objects A, B, C are given, we may obtain another unknown object D according to the analogical equation $A : B :: C : D$. In this example, the solution of the analogical equation is $D =$ "あなたはビールが好きですか。" (Do you like beer?). If we regard each sentence pair in a cluster as a pair $A : B$ (left to right or right to left), and any short sentence not belonging to the cluster as the object C, the analogical equation $A : B :: C : D$ of unknown D can be forged. Such analogical equation allows us to produce new candidate sentences by solving it.

4.2. *Experiments on new sentences generation and filtering*

For the experiment, we make use of the clusters we constructed in section 3.2 as rewriting models, the seed sentences as the input data for new sentences generation are the unique Chinese and Japanese short sentences from the 20,606 parallel sentences.

In this experiment, we generated new sentences with each pair of sentences in clusters for Chinese and Japanese respectively. It should be said that there may

exist no solution to an analogical equation, so that a new candidate is not coined each time. Conversely, each sentence pair in a cluster is a potential template for the production of new candidate sentences. This makes it possible to obtain more well-formed new sentences. We left aside clusters illustrating simple exchanges of digits, as they produce too many new sentences that are not really of interest. With this restriction, we generated about 62 million Chinese candidate sentences and more than 18 million Japanese candidate sentences. We extracted a sample of 1,000 sentences and checked their quality manually. The quality lies around 19% for Chinese and 50% for Japanese of correct sentences in syntax and meaning. Table 11 details the figures for this experiment.

Table 11. Statistics for new sentences generation in our experiments with Chinese and Japanese data.

		Chinese	**Japanese**
Initial data	# of seed sentences	16,259	20,079
	# of clusters	28,455	37,185
Generation	# of candidate sentences	61,565,221 Q=19%	18,403,787 Q=50%
Quality assessment	# of lines in references corpus	269,308	336,877
	# of new valid sentences (without begin/end marks)	25,212 N=6, Q=62%	26,003 N=7, Q=68%
	# of new valid sentences (with begin/end marks)	4,898 N=6, Q=99%	8,873 N=7, Q=99%

From the generated candidate sentences point of view, there are some characteristics in new sentences generation process. Table 12 and Table 13 give an example for new sentence generation with some seed sentences and a series of clusters in Chinese. Several important points that characterize the method are listed below:

- One seed sentence may produce different candidate sentences according to different clusters, because different clusters illustrate different linguistic features. For instance, the seed sentence 食物很不错。 'Food is very good.' in Table 13 produced 7 different candidate sentences according to all clusters given in Table 13.
- One seed sentence may produce different candidate sentences even for the same cluster as different sentence pairs (templets) are used. Table 13 illustrates this situation. For instance, the seed sentence 这个女孩长得美。 'The girl looks beautiful.' yielded two different candidate sentences using clusters identifier '3' which is examplified by '非常感谢 : 非常感谢作者' and '希望尽快解决 : 希望作者尽快解决';
- Different seed sentences may produce different candidate sentences according to the same cluster depending on the direction of the rewriting model, from

left to right or right to left. For instance, the seed sentences 经典电影 'classic movie' and 食物很不错。 'Food is very good.' produced four different candidate sentences according to cluster identifier '2' which represent the exchange of "经典" 'classic' to "很不错" 'very good' in one direction and the exchange of "很不错" 'very good' to "经典" 'classic' in the other direction.

- Different seed sentences may produce the same candidate sentence when passed to different clusters. For instance, the seed sentences 经典电影 'classic movie' and 糟糕电影 'bad movie' produced the same sentence 电影很不错 'the movie is very good' when passed to clusers '2' with the model of "经典" change to "很不错" and '4' with the model of "糟糕" 'bad' change to "很不错" 'very good' respectively.

Table 12. Examples of some seed sentences and a series of clusters in Chinese used for new sentence generation.

Seed short sentences	zh-clusters			Cluster identifier
	画面很美	:	画面很不错	
	挺美的	:	挺不错的	
	故事很美	:	故事很不错	
	真美	:	真不错	1
	美	:	不错	
必须感谢	美图	:	图不错	
经典电影	.....		.....	
这个女孩长得美。	经典啊	:	很不错啊	
糟糕电影	经典	:	很不错	
食物很不错。	经典故事	:	故事很不错	2
	喜欢经典	:	很不错喜欢	
	经典游戏	:	游戏很不错	
	.....		.....	
	非常感谢	:	非常感谢作者	
	希望尽快解决	:	希望作者尽快解决	3
	.....		.....	
	糟糕啊	:	很不错啊	
	糟糕游戏	:	游戏很不错	4
	.....		.....	

During the generation of candidate sentences, many invalid (e.g., "食物。经典") and grammatically incorrect (e.g., "这个女孩长得美作者") sentences are produced. To filter out these sentences and keep only well-formed sentences (e.g., "这个女孩长得不错。"), so as to ensure fluency of expression and adequacy of meaning, we eliminate any sentence that contains an N-sequence of a given length unseen in the reference corpus. Similar works using N-sequences to assess the quality of outputs of various NLP systems are works by C-Y. Lin and E. Hovy for summary generation[12], G. Doddington for machine translation[13], and also Y. Lepage and E. Denoual for

Table 13. The result of newly generated sentences according to the seed sentences and clusters in Table 12. The frequencies shows the times a candidate sentence has been generated using all possible clusters.

Seed short sentences		Newly generated sentences	Cluster identifier	Freq.
必须感谢	:	必须感谢作者	3	1
必须感谢	:	作者必须感谢	3	1
经典电影	:	很不错电影	2	4
经典电影	:	电影很不错	2	3
这个女孩长得美。	:	这个女孩长得不错。	1	4
这个女孩长得美。	:	这个女孩长得美作者。	3	1
这个女孩长得美。	:	作者这个女孩长得美。	3	1
糟糕电影	:	很不错电影	4	4
糟糕电影	:	电影很不错	4	3
食物很不错。	:	食物经典。	2	2
食物很不错。	:	食物。经典	2	1
食物很不错。	:	食物很美。	1	5
食物很不错。	:	美食物很。	1	1
食物很不错。	:	食物很不错作者。	3	1
食物很不错。	:	作者食物很不错。	3	1
食物很不错。	:	食物糟糕。	4	1

filtering paraphrases generation[14]. We tried to adequately change the length of the sequences of characters, so as to be able to obtain a satisfactory number of sentences with a high quality. Recall that 98% is the estimated quality of collected sentences data after filtering.

In a first attempt, we performed filtering directly based on all candidate sentences and the size of reference corpus in lines as given in Table 11 which also comes from the filtered sentences we collected from the Web, including the data for cluster construction and new sentences generation. The best quality we obtained with the values of length N=6 for Chinese and N=7 for Japanese were 62% and 68%. Among the Chinese and Japanese sentences kept, we found many incorrect sentences. We could easily identify that the problem laid with the beginning and the end of sentences. In English, the beginning and the end of sentences are well defined by the use of capital letters and the full stop. So as to reproduce similar conditions in Chinese and Japanese, we introduced begin/end markers to make sure that at least the beginning and the end of a sentence is correct. As a result, we obtained a lesser number of sentences that passed the test with 4,898 new valid sentences in Chinese and 8,873 new valid sentences in Japanese, using the values N=6 for Chinese and N=7 for Japanese. This time, however, the grammatical quality of these sentences, evaluated on sample of 1,000 sentences, was at least 99%. This means that 99% of the Chinese and Japanese sentences may be considered as grammatically correct (as shown in Table 11) when checked by native speakers.

Table 14 gives examples of newly generated sentences and the result of filtering using unseen N-sequences(N=6) for Chinese. For valid sentences, we remember their

corresponding seed sentences and the number of times they were generated by which cluster. The unacceptable new sentences are struck through in the table. The fourth newly generated sentence 价格高 'The price is high.' which is correct was filtered out due to its length of 5 (including begin and end markers) less than N=6. The size and coverage of reference corpus used is another factor that affects the number of kept valid sentences. For instance, the last sentence 各有各的特点,不能一概而论。 'Each one has its own characteristics, can not be generalized.' (correct) is also filtered out.

For valid sentences, we propose to deduce translation relations in two languages based on the translation relations between the seed sentences and the correspondence between the clusters that produced them.

Table 14. For each valid sentence, we remember its corresponding seed sentence and the number of times it was obtained (using different seed sentences or clusters), and the cluster that produced it.

Seed short sentences		Newly generated sentences	Cluster identifier	Freq.
我也非常喜欢音乐。	:	我也很喜欢音乐。	944	7
值得下载的游戏	:	值得推荐的游戏	608	10
值得推荐的软件	:	值得推荐的游戏	97	20
	:	~~价格高~~		
我最喜欢的这款软件很有意思。	:	我最喜欢的这款游戏很有市场。	1063	7
现在的版本比以前处理的快多了。	:	现在的版本已经比以前处理的快多了。	217	15
		~~乐器类卡片增加了的乐器~~		
		~~这个女孩长得。不错~~		
		~~小孩子很有教育意义。~~		
我最近很喜欢的一句话。	:	我最近非常喜欢的一句话。	944	7
		~~则改之。无则加勉~~		
		~~各有各的特点,不能一概而论。~~		

5. Deducing Translation Relations Between Newly Generated Sentences

In this paper, we examine an ideal configuration where the Chinese and Japanese seed sentences are parallel. Thus, for our experiment, the translation relations between the seed sentences rely on the 20,606 Chinese-Japanese parallel corpus. Then, we propose to extract corresponding clusters and compute the similarity between them so as to deduce and construct a Chinese-Japanese quasi-parallel sentences between new valid sentences.

5.1. *Extracting corresponding clusters by computing similarity*

Table 15 shows the flow of extraction of the changing characteristics in clusters. First, we extract the change between left and right sides in each cluster by finding

the longest common subsequence (LCS)[15] between each sentence pair. Then, we consider the changes between the left (S_{left}) and right (S_{right}) sides in one cluster as two sets. Finally, we perform the segmentation[k] for these changes in sets to obtain minimal sets of changes made up with words or characters.

Table 15. The flow of extraction of the changing characteristics cross the clusters in both languages.

	Chinese			**Japanese**		
Cluster	精美	:	画面很精美	綺麗	:	画面が[゛]とても綺麗
	可爱	:	画面很可爱	暗い	:	画面が[゛]とても暗い
	.....		.....	.....		.....
Changing extraction	⇓		⇓	⇓		⇓
	ε	:	画面很	ε	:	画面が[゛]とても
	ε	:	画面很	ε	:	画面が[゛]とても
	⇓		⇓	⇓		⇓
Sets of changes	(S_{left})		(S_{right})	(S_{left})		(S_{right})
	{ ε }	:	{画面很 }	{ ε }	:	{ 画面が[゛]とも}
Sets of changes	⇓		⇓	⇓		⇓
after segmention	{ ε }	:	{ 画面, 很 }	{ ε }	:	{ 画面, が[゛], とても}
Translation	⇓		⇓	⇓		⇓
and conversion	{ ε }	:	{ 画面 , 很 }	{ ε }	:	{ **画面**, が[゛], **很** }
Similarity computation		...			...	

Finding the corresponding clusters reduces to compute the similarity between two left sets (S_{left}) and two right sets (S_{right}) between Chinese and Japanese clusters. We make use of the EDR dictionary[l] and traditional-simplified Chinese conversion (Unicode Data-Traditional-Simplified-Variant[m]) and a Kanji-Hanzi Conversion Table[n] to translate all Japanese words into Chinese, or convert Japanese characters into simplified Chinese. We calculate the similarity between Chinese and Japanese word sets according to a classical Dice formula:

$$Sim = \frac{2 \times |S_{zh} \cap S_{ja}|}{|S_{zh}| + |S_{ja}|} \tag{2}$$

S_{zh} and S_{ja} denote the minimal sets of changes across the clusters (both on the left or right) in both languages (after translation and conversion). In our experiment, the formula for computing the similarity between Chinese and Japanese clusters is given in formula (3):

[k]Segmentation toolkits: Mecab, Part-of-Speech and Morphological Analyzer: URL: `http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html` for Japanese and Urheen, a Chinese lexical analysis toolkit (National Laboratory of Pattern Recognition, China) for Chinese.
[l]The EDR Electronic Dictionary: National Institute of Information and Communication Technology (NiCT). URL: `http://www2.nict.go.jp/out-promotion/techtransfer/EDR/index.html`
[m]`http://www.unicode.org/Public/UNIDATA/`
[n]`http://www.kishugiken.co.jp/cn/code10d.html`

$$Sim = \frac{1}{2}(Sim_{left} + Sim_{right}) \tag{3}$$

Table 16 gives some samples of the experiment results of found corresponding clusters between Chinese and Japanese. As the example shown in this Table, for the Chinese and Japanese clusters with the changes of "小说 : 电影很好看" and "小説 : いい映画", we obtained the Chinese translation for the Japanese word 映画 /eiga/ with 电影 /diànyǐng/ 'movie', and converted the Japanese kanji 説 /setsu/ in 小説 /shousetsu/ 'novel' to simplified Chinese hanzi 说 /shuō/ (here the Chinese word 小说 /xiǎoshuō/ also means 'novel'), thus, using this method, the similarity between these two clusters calculated by the formula (3) is 0.700. If all words and characters in the sets of changes (left and right) of a Japanese cluster could find their Chinese translations or can be converted into Chinese characters to match the sets of changes for a Chinese cluster (left and right), the similarity score will be 1.000. We also give these clusters with a similarity of 1.000 in Table 16. Such clusters are not necessarily clusters containing exact translations. For instance, the cluster: "新闻 : 采访" and "新聞 : インタビュー" obtained a similarity of 1.000 because of the translation relation between "采访" and "インタビュー", and kanji-hanzi conversion of Japanese word "新聞" to simplified Chinese word"新闻". However, actually the meaning of Japanese "新聞" /sinbun/ 'newspaper' is different from Chinese "新闻" /xīnwén/ 'news'.

Table 16. Examples of corresponding clusters (with the changes between the left and right sides) with high similarity scores.

Chinese Cluster			**zh ⇐ ja**	**Japanese Cluster**			**Similarity**
面包	:	日本料理		パン	:	日本料理	1.000
喜欢	:	讨厌		好き	:	嫌い	1.000
新闻	:	**采访**		**新聞**	:	**インタビュー**	1.000
很	:	非常	EDR dictionary	超	:	とても	1.000
但是	:	ε	+	でも	:	ε	1.000
照片	:	ε	TS conv.	写真	:	ε	1.000
ε	:	她	+	ε	:	彼女	1.000
ε	:	非常	kanji-hanzi conv.	ε	:	非常に	0.833
小说	:	**电影很好看**		**小説**	:	**いい映画**	**0.700**
十分	:	非常		ε	:	とても	0.500

$$Sim = \frac{1}{2}\left(\frac{2 \times |\{小说\}|}{|\{小说\}| + |\{小说\}|} + \frac{2 \times |\{电影\}|}{|\{电影, 很, 好看\}| + |\{いい, 电影\}|}\right) = \frac{1}{2}\left(1 + \frac{2}{5}\right) = 0.700$$

5.2. *Experiments and results*

The proposed method described above was applied for deducing translation relations between valid newly generated sentences in an actual experiment based on the following data files:

- 20,606 Chinese-Japanese parallel corpus with translation similarities between the seed sentences for new sentences generation as presented in section 2.
- 4,898 valid newly generated Chinese sentences (8,873 for Japanese) with their corresponding seed sentences, identifiers of the clusters that they generated from, the data obtained as reported in section 4.2.
- Extracted corresponding clusters obtained as described in section 5.1 across languages with the sets of changes in both languages, identifiers and the similarity scores.

As a final result,Table 17 gives the statistics for the obtained quasi-parallel Chinese-Japanese sentences deduced by computing the similarity between the clusters across languages and the similarity between sentences for new sentence generation.

Table 17. Statistics for the obtained quasi-parallel new sentences in Chinese and Japanese by setting different threshold value of Dice.

$\mathbf{Sim_{C_{zh}-C_{ja}}}$	**# of corresponding clusters**	**Extracted quasi-parallel**	**Ture translation**	
≥ 0.3	6,030	1,837	1,124	61.2%
≥ 0.4	4,192	1,531	970	63.3%
≥ 0.5	2,299	1,260	887	70.4%
≥ 0.6	1,693	1,247	878	70.4%
≥ 0.7	1,676	1,244	878	70.6%
≥ 0.8	1,229	1,240	877	70.7%

The final data, i.e., the quasi-parallel corpus is presented in the format shown in Table 18, which shows a sample of the data obtained with our proposed method. N_{zh} and N_{ja} denote the valid newly generated sentences, Sim_{zh-ja} is the similarity between the seed sentences, and $Sim_{C_{zh}-C_{ja}}$ is the similarity between the corresponding clusters from which these two new sentences were produced, $F(zh)$ and $F(ja)$ denote the number of times these new sentences were generated with the two corresponding clusters.

Table 18. Samples of deduction result with their associated similarity scores between newly generated sentences in two languages. Clusters cross languages as insertion, deletion or substitution rewriting models for these new sentences generation. The words or characters in brackets are the deletion parts of these new sentences, in parentheses are the insertion parts of these new sentences.

Quasi parallel Chinese-Japaense sentences (new)		**Features**			
N_{zh}	N_{ja}	Sim_{zh-ja}	$Sim_{C_{zh}-C_{ja}}$	$F(zh)$	$F(ja)$
历史上遗留下来的边界[问题]	歴史上残されてきた国境[問題]	1.000	1.000	10	7
根据同名小说摄制的[电影]	同名の小説に基づいて制作した[映画]	1.000	1.000	21	13
我觉得那是昭和初期的[电影](作品)。	それは昭和初期の[映画](作品)だと思います	1.000	1.000	25	17
[雨]水从山坡上哗哗地冲泻下来了。	[雨]水が山からざあざあと流れてきた。	1.000	1.000	5	8
明天吃[面包](日本料理)	明日は[パン](日本料理)にする	1.000	1.000	9	8
[一切]都是按照计划进行的。	[すべて]は計画とおりにいった。	1.000	1.000	9	8
新学期[开始]时是书店的旺月	新学期の[はじめ][は]本屋の書き入れ時だ	1.000	0.833	9	6
没有恢复希望的[病人][。]	回復の望みがない[病人]。	1.000	0.833	5	11
(你喜欢)[想]喝咖啡(吗)	(あなたは)コーヒーが[飲みたい](好きですか)	1.000	0.708	5	7
需要紧急处理的[问题](事)	もっとも緊急を要する[問題]	1.000	0.500	8	4
(那)(是)当然了	(それ)(は)当たり前[です]	1.000	0.500	17	15
街上人来人往热闹[非常](极了)	通りでは人の往来が激しく[非常に]にぎやかだ	1.000	0.333	18	9

6. Conclusion

We presented a technique which uses analogical associations to construct a free Chinese-Japanese quasi-parallel corpus based on sentences that we collected from the Web with the concern of avoiding any copyright problem. From 47,674 sentences in Chinese and 95,130 sentences in Japanese, we constructed 28,455 analogical clusters in Chinese and 37,185 in Japanese. Different clusters illustrate different linguistic features. These clusters served as rewriting models to generate new sentences. We actually obtained a very large amount of new candidate sentences. To ensure fluency of expression and adequacy of meaning we filtered the generated sentences by the N-sequences method combined with the introduction of begin/end markers. 4,898 Chinese new sentences and 8,873 Japanese new sentences were obtained after filtering. Their grammaticality and semantic validity was evaluated by sampling and was found to be of at least 99% for both Chinese and Japanese. This quality is of the same level as the quality of the resources we started from. Such valid new sentences are not necessarily the paraphrases in comparison with the seed sentences. We then deduced the translation relations between valid newly generated short sentences between Chinese and Japanese, based on the similarity between the seed sentences and the clusters they were generated from.

In this paper, we examined an ideal configuration where the Chinese and Japanese seed sentences were parallel sentences, We automatically obtained 1,837 quasi-parallel new generated sentences without copyright problems as all sentences have been created by our method. The result is a resource of pairs of sentences in Chinese and Japanese with associated similarity translation scores.

We are continuously collecting short sentences from the Web, and continuously construct new clusters with the new different filtered sentential data. We merge these clusters according to the set of exchanges of each cluster, and aim to obtain larger sizes of clusters for new sentences generation. Recently we obtained 78,630 Chinese clusters and 137,231 Japanese clusters in this way. With the same number of seed sentences as input data, we obtained 22,313 Chinese new sentences and 30,906 Japanese new sentences after filtering by N-sequences with the larger size of reference corpora (966,457 for Chinese and 975,518 for Japanese) that remain to align for translation similarity.

In future work we propose to extract more quasi Chinese-Japanese parallel sentences based on the method we described in this paper, even from comparable resources and make them freely available to the community to foster research in Chinese-Japanese machine translation.

Acknowledgments

This work was supported in part by Foreign Joint Project funds from Kitakyushu Foundation for the Advancement of Industry, Science and Technology (FAIS).

References

1. P. Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation, In *Proceedings of the tenth Machine Translation Summit (MT Summit X)*, Phuket, Thailand, pp. 79-86, Sept. 2005.
2. R. Steinberger, A. Eisele, S. Klocek, S. Pilos, and P. Schlüter. DGT-TM: A freely Available Translation Memory in 22 languages, In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pp. 454-459, 2012.
3. Y. Zhang, K. Uchimoto., Q. Ma and H. Isahara. Building an Annotated Japanese-Chinese Parallel Corpus—A Part of NICT Multilingual Corpora, In *Proceedings of the Tenth Machine Translation Summit (MT Summit X)*, pp. 71-78, 2005b.
4. M-Y. Yang, H-F. Jiang, T-J. Zhao and S. Li. Construct Trilingual Parallel Corpus on Demand, In *Proceedings of Chinese Spoken Language Processing, 5th International Symposium, ISCSLP 2006*, Volume **4274** of Lecture Notes in Computer Science, pp. 760-767, Singapore, December 13-16, 2006.
5. Y. Lepage and E. Denoual. Purest ever example-based machine translation: detailed presentation and assessment, *Machine Translation*, **19**, pp. 251-282, 2005b.
6. P. D. Turney and M. L. Littman. Corpus-based Learning of Analogies and Semantic Relations, *Machine Learning*, **60**(1-3), pp. 251-278, 2005.
7. N. Stroppa and F. Yvon. An analogical learner for morphological analysis, In *Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL 2005)*, pp. 120-127, Ann Arbor, MI, 2005.
8. J-F. Lavallée and P. Langlais. Morphological acquisition by formal analogy, In *Morpho Challenge 2009*, Corfu, Greece, Oct. 2009.
9. Y. Lepage. Solving analogies on words: an algorithm, In *Proceedings of the 36th Annual Conference of the Association Proceedings of the 36th Annual Conference of the Association for Computational Linguistics (COLING-ACL'98)*, Volume I, pp. 728-735, Montréal, Aug. 1998.
10. L. Allison and T. I. Dix. A bit string longest common subsequence algorithm, *Information Processing Letter*, **23**, pp. 305-310, 1986.
11. F. de Saussure. Cours de linguistique générale, Payot, Lausanne et Paris, [1ère éd. 1916] edition, 1995.
12. C-Y. Lin and E. Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics, In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, HLT-NAACL '03*, pp. 71-78, 2003.
13. G. Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics, In *Proceedings of the Human Language Technology Conference (HLT-2002)*, pp. 128-132, San Diego, CA, USA. Morgan Kaufmann, 2002.
14. Y. Lepage and E. Denoual. Automatic generation of paraphrases to be used as translation references in objective evaluation measures of machine translation, In *the 3rd International Workshop on Paraphrasing (IWP2005)*, pp. 57-64, 2005a.
15. R. A. Wagner and M. J. Fischer. The string-to-string correction problem, *Journal of the Association for Computing Machinery*, **21**, pp. 168-173, 1974.
16. Y. Lepage and C-L. Goh. Towards automatic acquisition of linguistic features, In *Proceedings of the 17th Nordic Conference on Computational Linguistics (NODALIDA 2009)*, pp. 118-125, Odense, May, 2009.
17. Y. Lepage. Analogy and formal languages, *Electronic notes in theoretical computer science*, **53**, pp. 180-191, April 2004a.
18. Y. Lepage. Lower and higher estimates of the number of "true analogies" between sentences contained in a large multilingual corpus, In *Proceedings of the 20th international conference on Computational Linguistics (COLING '04)*, volume 1, pp. 736-742, Cenève, August, 2004b.
19. A. Delhay and L. Miclet. Analogical equations in sequences: Definition and resolution, *Lecture Notes in Computer Science*, 3264, pp. 127-138, 2004.
20. N. Stroppa and F. Yvon. An analogical learner for morphological analysis, In *Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL 2005)*, pp. 120-127, Ann Arbor, MI, June, 2005.

Wei Yang

She received the Master Degree in 2012 from Waseda University, graduate school of Information, Production and Systems. During her Master Course, her research interests in combining several automatic techniques to build Chinese-Japanese lexicon from freely available resources and make them free available for users and researchers in Natural Language Processing and Machine Translation. She is currently a Ph.D. candidate at the Waseda University, graduate school of Information, Production and Systems. Her research interests are in Natural Language Processing, Machine Translation, especially between Chinese and Japanese.

Hao Wang

He is currently a master student at the Waseda University, graduate school of Information, Production and Systems. His research interests include Natural Language Processing, Data Mining, Corpora Construction and Example-based Machine Translation.

Yves Lepage

He received his D.E.A. and Ph.D. degrees from Grenoble university, France, in GETA under the supervision of Prof. Vauquois and Prof. Boitet. After a post-doctorate at ELSAP, university of Caen and EDF, Paris, he joined ATR labs, Japan, where he worked as an invited researcher and a senior researcher until 2006. In 2003 he got the habilitation for his habilitation thesis on proportional analogies in linguistics. In February 2006, he got the qualification for full professorship from the National Board of French Universities in both linguistics and computer science and became full professor at the University of Caen Basse-Normandie in October 2006. He joined Waseda University, graduate school of Information, Production and Systems in April 2010. His research interests are in Natural Language Processing, Machine Translation, and in particular Example-Based Machine Translation. He is a member of the Japanese Natural Language Processing Association. He is a member of the board of the French Natural Language Processing Association, ATALA, and editor-in-chief of the French journal on Natural Language Processing, TAL.